ARI TECHNICAL REPORT
TR-79-A8

Principles of Work Sample Testing:
I. A Non-Empirical Taxonomy of Test Uses

LEVEL II

by

Robert M. Guion

BOWLING GREEN STATE UNIVERSITY
Bowling Green, Ohio 43403

April 1979

Contract DAHC 19-77-C-0007

Prepared for

U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
5001 Eisenhower Avenue
Alexandria, Virginia 22333

U. S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

JOSEPH ZEIDNER
Technical Director

WILLIAM L. HAUSER
Colonel, US Army
Commander

NOTICES

C-2

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

19

1. REPORT NUMBER: TR-79-A8

2. GOVT ACCESSION NO.: 18 ARI

3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle):

6 PRINCIPLES OF WORK SAMPLE TESTING.
I. A NON-EMPIRICAL TAXONOMY OF TEST USES.

5. TYPE OF REPORT & PERIOD COVERED: 9 Final rept. 15 Nov [redacted]76 - 15 Jun[redacted]78

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): 10 Robert M. Guion

8. CONTRACT OR GRANT NUMBER(s): 15 DAHC19-77-C-0007

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Bowling Green State University
Bowling Green, Ohio 43403

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 16 2Q161102B74F

11. CONTROLLING OFFICE NAME AND ADDRESS:
U.S. Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, Virginia 22333

12. REPORT DATE: 11 Apr[redacted]79

13. NUMBER OF PAGES: 49

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): 12 60p.

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Monitored by G. Gary Boycan, Engagement Simulation Technical Area, Army Research Institute.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Measurement theory, psychometrics, work sample testing, validity, content-referenced testing, criterion-referenced testing, latent trait theory, generalizability theory

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Challenges to classical psychometric theory are examined in the context of a broader range of fundamental, derived, and intuitive measurements in psychology; the challenges include content-referenced testing, latent trait theory, and generalizability. Psychological measurement is classified according to purpose, setting, variables, and methods of measurement. The challenges and the classifications are examined for special implications for work sample testing.

(Continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

061850

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. (continued)

The report, the first of four, is written for psychologists interested in psychometrics.

Accession For
NTIS GRA&I ☑
DDC TAB ☐
Unannounced ☐
Justification ______

By ______
Distribution/
Availability Codes
Dist. | Avail and/or special
A |

ACKNOWLEDGMENTS

Conversations with many people have influenced the content of these reports, but the influence of a few of these has gone well beyond that of the usual conversation and should be acknowledged. Examples include Fumiko Samejima and Frederic Lord, who introduced me to the mysteries of latent trait theory, Robert Ebel, whose pithy comments on the obfuscation inherent in discussions of validity have inspired much of the questioning behind these reports, and Sam Messick and Mary Tenopyr, who convinced me to break the content validity habit. The contributions of these people, although unintentional, are gratefully acknowledged.

Special acknowledgment is due to the help of Robert Brennan, of American College Testing, for graciously helping me come to some insights about generalizability theory, and to Gail Ironson, who has been a colleague and regular consultant on matters of mathematical and logical foundations in the three areas of challenge (content-referenced testing, latent trait theory, and generalizability theory) around which these reports have been written. Without her advice and help, and her availability for trying out ideas, they might not have been written; she shares authorship, in fact, on the paper on generalizability.

Most of all, I want to express thanks to Gary Boycan of the Army Research Institute, whose initiative has made possible the opportunity to read, sit, stare, and think about matters that have long troubled me.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

PRINCIPLES OF WORK SAMPLE TESTING: I. A NON-EMPIRICAL TAXONOMY OF TEST USES

BRIEF

Because classical psychometric theory often seems inadequate for the development and evaluation of work sample tests, and because recent challenges to classical theory have had promising implications, the conceptual foundations of work sample testing need to be examined and clarified. This report, the first in a series of four, attempts to provide a background for that examination by considering the full scope of measurement in psychology. The purpose is to determine whether different kinds of measurement, or different circumstances of measurement, have different implications for the development and evaluation of measurement procedures.

The most fundamental approach to measurement is mathematically formal; it conforms to certain mathematically stated axioms, principally the axiom of transitivity. Fundamental measurement is expressed in formally defined units which are widely accepted throughout the scientific community. The use of such formal measurement provides rather direct descriptions, with little or no need for inferences, of the attributes of objects being measured. An example of such measurement is linear distance. One does not speak of "inferring" the length of an object through measurement, although it would be true, because the inference and the fact of the measurement are very nearly the same.

Most measurement in psychological research, and particularly the measurement described by classical psychometric theory, provides only signs from which inferences are drawn about the attributes of interest. The unit of measurement is typically the standard deviation of the distribution of a set of measurements, not a mathematically defined formal unit; traditional psychometric measurement is said, therefore, to be "norm-referenced." That is, the meaning of a score is defined relative to its position within the distribution of scores; in contrast, fundamental measurement can be applied to the single case, defining the meaning of a "score" in terms of the units of measurement in the scale used.

Three challenges to classical psychometric theory have gained in attention in recent years. One of these is a trend toward greater preference for content-referenced measurement as distinguished from norm-referenced measurement. Another is latent trait theory, which provides an analog, at least, to a mathematically formal unit of measurement. The third is generalizability theory, which seeks a more

precise understanding of the errors of measurement. All of these challenges seem to have special significance for work sample testing.

To provide a framework within which to consider classical psychometric theory and these challenges, a tentative taxonomy of psychological measurement is proposed. With it, the special issues in work sample testing can be viewed in the larger context of measurement in psychology generally. Four specific taxonomies are proposed: classifications of (a) the purposes of measurement, (b) settings in which measurements are obtained, (c) variables or attributes to be measured, and (d) the methods of measurement in psychology.

Six broad purposes of measurement are identified:

1. Evaluation of materiel, processes, or programs to permit organizational decisions to be made about them.

2. Organizational trouble shooting to identify needs for corrective actions concerning personnel units.

3. Individual diagnosis identifying strengths and weaknesses of individuals, either internally or relative to others.

4. Certification of individual proficiency or need, or levels of these, such as in the skill qualification testing program.

5. Prediction of future performance or characteristics of individuals, such as prediction for selection decisions.

6. Evaluation of other measurements, such as the use of one measurement as a criterion in the validation of another one.

Three types of measurement settings are defined. Types of variables are presented under two subheadings, attributes of people and attributes of tasks. Seven categories of the personal attributes are listed in decreasing order of objectivity of measurement, and a similar order is tentatively proposed for nine categories of task variables. Five kinds of measurement methods are identified, again in decreasing order of probable objectivity in measurement, ranging from the use of special instrumentation to the use of ratings.

Most purposes of measurement require, at least for the evaluation of measurements, at least the potential for substantial variance; argument that mastery testing, for example, should have low variance is rejected. Regardless of purpose, some form of generalizability is needed, although the diagnostic and certification purposes emphasize

the generalizability of scores while prediction requires generalizability of relationships. Regarding the categories of settings, the same statement is appropriate: generalizability across settings, either of scores or of relationships, seems universally necessary.

The implications of the joint classification of variables and of the methods for measuring them provide more diverse implications. For the most highly objective combinations, measurement must be accurate and interpretable in relation to a standard. Since work sample tests strive for objectivity, the same implications exist for them. The more subjective combinations require research into the acceptability of possible inferences as the principal form of evaluation.

INTRODUCTION

The well-established technology for aptitude testing seems inadequate for some purposes, including certification testing by work samples. In recent years, challenges to classical psychometric theory have come from many sources (e.g., Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Lord, 1952; Muir, 1977; Popham & Husek, 1969). This report, and the three that follow, will consider both classical theory and its challenges in examining (a) some special problems of work sample testing, (b) some relatively new developments in measurement, and (c) some old measurement ideas that are often ignored in psychometric discussions. By examining and clarifying the conceptual foundations of work sample testing, these papers will offer principles for the construction, use, interpretation, and evaluation of work sample tests in the broader context of general problems in the measurement of psychological variables.

The present report will identify the place of work sample testing in the context of a non-empirical taxonomy of general psychological measurement. The taxonomy will be described, and its implications for test evaluation will be presented with special emphasis on work sample testing. The second paper looks broadly at the scope of systems for evaluation of personnel testing programs. Evaluation includes psychometric concepts of validity, but it is not restricted to them.

With these broad perspectives as context, the third paper will focus explicitly on the construction and validation of work sample tests. Since the principal requirement to satisfy in work sample testing is generalizability of scores, the final paper in the series will be concerned explicitly with the problems and opportunities of different kinds of generalizability research for work samples and work sample validities.

A SYNOPSIS OF MEASUREMENT THEORY

Measurement is a characteristic scientific endeavor. No field of scientific enterprise can progress far without operationally defining, classifying, and quantifying its variables. Applied science relies especially heavily on the quantification of its subject matter. Measurement is not unique to psychological research, nor is preoccupation with an underlying theory of measurement a special prerogative of psychometrics.

Fundamental to any discussion of measurement is the fact that one does not measure objects or people; rather, one measures attributes of objects or people. Measurement implies the assignment of numbers to represent attributes according to some specified set of rules. Systems for assigning numbers can be devised for representing the weight of objects, the amount of information in a message, the amount of perceptual skill characterizing an individual, or the quality of an individual's performance. An acceptable system of measurement assigns numbers to represent only one attribute; other numbers, assigned according to other rules, can represent other attributes of the same objects, messages, people, or performance.

KINDS OF MEASUREMENT

It is useful to distinguish different kinds of measurement. Some approaches to measurement are so constructed that the numerical result in measuring an attribute of something is understood primarily with reference to the measurement system itself. In such measurement systems, the rules for assigning numbers to represent quantitites are so definite, unambiguous, and widely accepted that an obtained number has an immediate and obvious descriptive meaning. Many of these intrinsically

obvious measurement processes have a foundation in clearly established natural law. Torgerson (1958) referred to such measurement as fundamental. The best examples of such fundamental measurement are physical measurements such as counting objects, weighing objects, measuring distances, and the like.

He also described derived measures, those which are derived from fundamental measurement with a similar kind of internally consistent meaning. Examples include more complex kinds of physical measurement, such as the measurement of density as a ratio of mass to volume. While these may not take their meaning in a wholly internal way, as in more nearly fundamental measurements, they take their meaning in the relationships of established scientific law relating an attribute to other attributes.

Both kinds of measurement described above are mathematically formal systems; that is, they conform to certain basic mathematical axioms such as those of transitivity or additivity, and that conformity can be demonstrated through formal mathematical proofs.

Although the most obvious examples of mathematically formal measurement are physical measurements, psychology is not without such formal systems of its own. Quite apart from the obvious behavior frequency counts (which are, of course, physical measurements of rate of occurrence), psychology has specialized fields of mathematical measurement theory such as information theory and signal detection theory. These approach measurement formally with neither interest in nor need for the conventional psychometric theory developed for traditional mental testing.

In contrast, other measurement derives meaning inferentially

more than directly descriptively. For example, formal physical measurements may be used to describe directly an attribute from which some other attribute is inferred; we speak (perhaps erroneously) of having "measured" the inferred attribute. An excellent example is the galvanic skin response; literally, one measures electrical resistance on the surface of the skin, but changes in that resistance are used for inferring changes in emotionality, and GSR is said to be a measure of emotion.

Much of psychological measurement is derived measurement, but it is statistically derived. It is not derived from statements of invariant lawfulness, as in the measurement of density; it is formally derived from statistical analyses and assumptions. The best examples of statistically derived measurement in psychology are those stemming from research in psychophysics, such as using Thurstone's Law of Comparative Judgment (Thurstone, 1959). The early history of mental testing proceeded in an analogous way; each item in a test was treated as a stimulus item, the response to which had some probability of providing an appropriate inference. The probability of an appropriate inference was increased by repeated stimulation, i.e., by using several items to make up a total measure or score. Modern computerized adaptive or tailored testing is a further example of statistically derived measurement, differing from earlier testing more in mathematical sophistication than in principle.

Statistically derived measurement is no less formal, and no less rigorous, than mathematically formal measurement derived from fundamental measurements. Most statistically derived psychological measurement has its own unique "mental unit of measurement" (Thurstone, 1959, p. 50). Whether it is the discriminal dispersion of judgments of scale separation, the variance in a set of test scores, or a

hypothetical scale for measuring latent ability, the unit of measurement in most mental measurement is the standard deviation.

Inferences may also be drawn from less formally developed measuring instruments. A fourth category of the kinds of measurement includes what can best be described as intuitive measurements. Many index numbers are established by intuitively combining a host of considerations; ad hoc tests may be constructed without prior statistical analysis but with some degree of rational thought; perhaps the best example of intuitive measurement is the ubiquitous five-point rating scale which is applied willy-nilly, without any formalisms or supporting data. Intuitive measures can be highly useful. Much of economic theory has been developed using such index numbers. As research progresses with such measurement schemes, lawful relationships are often identified which permit the development of more formal approaches to the measurement of the same variables.

MEASUREMENT OF WORK SAMPLE PERFORMANCE

Work sample testing may use all of the kinds of measurement in measuring attributes of either the work process or of the product (Shimberg, Esser, & Kruger, 1972). Intuitive scales may be used to rate or evaluate the process. Performance might be scored like paper-and-pencil tests are scored (which some forms of work sample tests actually are), using the theoretical foundations and principles for selecting items and evaluating scores used in traditional test construction and evaluation. Fundamental measures may be used to describe the product or result of performance; quality of performance can be inferred by weighing, by determining a physical breaking point, measuring conformity to tolerances, or by using other forms of fundamental, physical measurement of chosen attributes of a physical product.

Classical psychometric theory, which is but one theory among many, does not traditionally apply to, and may be inadequate for, some kinds of work sample measurement. Much mischief and confusion can result from misguided attempts to squeeze work sample testing into the same rubric used for the evaluation of inferences drawn from aptitude tests, even though many work sample variables can be appropriately handled within a conventional psychometric theory.

In considering alternatives for the evaluation of performance on work samples, it is instructive to consider challenges to traditional theory that have been offered in recent years. Perhaps the most active field of challenge is that known as content-referenced measurement, among other names, with its insistence that performance be measured not in terms of standard deviations from a sample or population mean but in terms of reaching or deviating from a specified standard level of performance (Glaser & Klaus, 1962).

Another emerging challenge to traditional psychometric theory comes from latent trait theory (Lord, 1952), or latent structure analysis (Lazarsfeld, 1950), which attempts to identify item characteristics as essentially sample-free estimates of item parameters instead of item statistics based on the sample at hand. Characteristics of a test can then be defined in terms of the characteristics of independent items comprising the test.

A third challenge comes from generalizability theory (Cronbach et al., 1972), which questions the adequacy of the traditional true score and error score division of obtained scores; it works instead to allocate the portions of total obtained score performance among various facets or conditions of measurement. In short, generalizability theory argues that it is the consistency or dependability of

measurement over varying conditions that is the important point in the evaluation of measurement.

These challenges are all relevant to the development and evaluation of work sample tests. For example, if a particular work sample is devised for a welder, and if all of the people who are administered the test perform poorly on it, there is little benefit to be derived from identifying certain people as having performed better than others; the significant statement is the content-referenced interpretation that they all performed below standard. Since work is rarely conducted under well-controlled, standard conditions, the stability of work sample performance across a reasonable range of circumstances is certainly important. The applications of latent trait theory are perhaps less obvious; it is sufficient here to note that such applications can provide a basis for standardizing interpretations of content-referenced tests over different samples of people tested in different locations or at different times.

In short, these challenges to traditional psychometric theory, and perhaps others, may lead to a newer and firmer foundation for work sample testing. It is therfore useful to examine work sample testing in context in the gamut of psychological measurement. The purpose of this examination is to determine whether different kinds of measurement, or measurement in different circumstances, have different implications for the development and evaluation of measurement procedures, particularly by work sample testing.

CONSIDERATIONS FOR A MEASUREMENT TAXONOMY

Psychological measurement does not occur as a disembodied abstraction. It occurs in the context of a broader purpose than measurement

per se, and it occurs within a broader environmental context. Purpose and setting, perhaps as much as the measurer's skill, determine what is to be measured and how one may go about it. The purposes, settings, variables, and techniques define a "gamut of psychological measurement" much more extensive than is ordinarily considered. The principal purposes of this report are (a) to suggest ways in which each of these may be classified and (b) to suggest implications of these classifications for the development and evaluation of specific approaches to measurement.

Personnel testing -- indeed, the testing movement as a whole -- occupies a relatively small portion of the total field of psychological measurement. Work sample testing, even broadly defined, occupies a correspondingly small place in the personnel testing domain. The tunnel vision of overspecialized theorizing can and does permit competent theory and practice in that branch of measurement traditionally known as psychometrics, but test theory and practice can be enriched by taking cues from a broader vision of measurement.

The implications of the different categories can sometimes focus on some kinds of descriptions of appropriate measurement, descriptions that can be expressed as simple dichotomies. The introductory remarks have emphasized one of these, the distinction between fundamental, descriptive measurement internally interpretable and more nearly intuitive, inferential measurement. Classical psychometric theory emphasizes the latter. It has also been pointed out that another possible dichotomous classification distinguishes norm-referenced from content-referenced measurement; classical theory addresses the former. Another possible dichotomy distinguishes measures of maximum performance from measures of typical performance; classical theory addresses both so long as performance can be inferred normatively.

In the sections that follow, purposes, settings, variables, and methods in measurement will be further divided, quite arbitrarily, into a number of categories. The categories are not exhaustive or particularly fine; they have not been empirically identified, nor has any attempt been made to ascertain their usefulness by determining empirically the reliability with which they can be used to classify actual measurement programs. They can, nevertheless, provide some insights into the place occupied by personnel testing, and especially by work sample testing, in the broader scheme of psychological measurement; they may also suggest principles for the evaluation of specific kinds of work sample measurement.

PURPOSES IN PERSONNEL TESTING

Nearly every use of personnel tests has in some sense a unique purpose, and each purpose has its own implications for the development and evaluation of measuring procedures. Nevertheless, some broad classes of reasonably similar purposes may be identified and examined for their special kinds of implications.

<u>Evaluation of Materiel, Processes, or Programs</u>. One purpose of personnel testing is to provide a dependent variable. Hypotheses that particular equipment or procedures or programs either will improve performance of personnel and should be adopted, or will have no effect or a negative effect on performance and should not be adopted, are tested in decision-oriented research. For an example, see Dobbins and Kendrick (1965) on the use of lenses in personnel detection within tropical forests.

In such circumstances, the psychological measurement of interest is usually a measure of performance. There are many ways to assess

performance; examples include ratings, counts of production or other achievements, output/input ratios of various kinds or records of production, or personnel problems over a period of time.

All of these imply some sort of work sample for program evaluation. The term is being used terribly broadly here to make the point; proficiency ratings, for example, are typically assessment of performance sampling a specified period of time, hence of a work sample of sorts. The question in assessing performance in these studies is not whether a sample of work is to be observed and evaluated but rather how effectively the sample of performance can be assessed. The first question in evaluating performance measurement is whether the sample of performance observed in the experimental setting is representative of performance in real or typical or targeted circumstances. There are also basic questions of (a) whether the performance is directly observed or only vaguely perceived (as in supervisory ratings) and (b) whether the numbers representing evaluations of performance in fact reflect irrelevant attributes of either the behavior, the worker, or the observer.

The measurement of performance in the experimental situations typical of these studies is rarely concerned with individual differences. The important unit of analysis is the group, not the individual, and the typical measure of interest is the mean performance of various experimental or control groups; "validity" is expressed as the significance of differences between these mean levels of performance. Occasionally the variance of subgroup performance will be the statistic of interest. Very rarely is the individual measure the measurement of concern in these experimental circumstances. Individual differences are usually (although improperly) treated as error variance. The reason, of course, is that the purpose of the

research is to make a decision about organizational practices or procedures, not a decision about individuals.

Organizational Trouble Shooting. A potential but not well explored use of personnel measurement is for the diagnosis or identification of organizational problems (Boyd, 1961). Measures of job satisfaction may be taken in different aspects of an organization to try to identify subgroups who may be pockets of discontent. Job knowledge tests could be given in different units to identify similar pockets of ignorance. Psychological assessment techniques may be used to identify areas of inefficiency, of inappropriate behavior, or of personnel misclassification. Most such studies are correlational in nature; such studies should attempt to maximize the relevant variances among individuals, somewhat like a magnifying glass. Other attempts to diagnose organizational problems may use quasi-experimental designs; in these studies variance within groups may be treated as error to be minimized while seeking to maximize between-group differences.

Individual Diagnosis. The term diagnosis is not restricted to clinical use. In many personnel testing uses, the purpose is to identify individual strengths and weaknesses. Sometimes the intent is to identify a person's own relative strengths and weaknesses, regardless of level. In other cases, one asks whether one individual measures up well or poorly in relation to others or, perhaps, to some standard on any given attribute. These are inferential measures; they should be chosen or constructed to yield the most acceptable and useful descriptions of the attributes assessed with minimal contamination from other attributes. A critically important issue in making comparisons is whether the measurements of different variables or from different samples can be expressed in a common metric.

Certification. A common purpose of measurement is to certify to decision-makers that individuals have levels of attributes appropriate to specific decisions. A high score on a licensing examination tells the Board of Examiners that it can decide to certify to the public that the person is competent or has certain knowledge essential to competence. The Army system of skill qualification testing is another example (Maier, Young, & Hirshfeld, 1976). Certification does not necessarily indicate anything desirable; a clinical psychologist may be required, for example, to certify to the court that a particular person is incompetent to stand trial, or to participate in his own defense, or some other form of incompetence. In personnel measurement, certification usually is intended to assure decision-makers that certain individuals have (or do not have) certain qualifications necessary for effective performance.

Certification usually implies a dichotomous decision. An individual will either be accepted for a job or for training or will not be accepted; measurement can likewise be reduced to a simple dichotomy. It should not be believed, however, that dichotomous scoring eliminates variance among people chosen; variance, like the poor, will be with us always. What is implied is that, for some uses of measurement, the amount of variance within a group may seem trivial. Measurement for certification may, therefore, be considered similar to measurement for organizational decisions or for trouble shooting; the problem may be to minimize within-group variance and maximize between-group variances.

Prediction of Future Status Events or Performance. All of the preceding categories logically imply a sort of prediction. There are, however, many purposes which may be explicitly stated in formal language as predictive hypotheses.

Where prediction is the explicit purpose, two or more measurements are involved: the measurement of the future variable -- individual status or performance or the occurrence of an event -- and the measurement of the predictor. The time element is an important part of the predictive hypothesis, and the evaluation of measurement may include an evaluation of the appropriateness of the elapsed time or other circumstances under which the measurements are taken. Descriptive measurements both at the time of prediction and the future time need evaluation. Most important is the need to evaluate not only the measurement but the tenability of the hypothesis itself.

There is nearly an infinite variety of things to predict in personnel testing. One may wish to predict whether training will be completed, level of proficiency at the conclusion of training, or proficiency or other forms of behavior at some stabilizing period after training has been completed. Each of these may call for slightly different evaluations of measurement. If one attempts to measure proficiency at the end of training, the measurement may seek to assess maximum performance capability with reference to some standard. Depending on the specific hypothesis, prediction of on-the-job proficiency may require measurement of either typical or maximum performance.

<u>Evaluation of Other Measurement</u>. To complete the list of purposes, it is necessary to point out that some personnel assessment is done primarily in the validation of other measurement. It may serve as a criterion measurement, as in prediction of future performance, or as the measurement of a hypothesis tested in the evaluation of construct validity.

TYPES OF MEASUREMENT SETTINGS

There is almost an infinite variety of situations in which measurements are taken. Each category below could be subdivided, some of them many times, with an increase in the precision with which settings can be described. A relatively small number of categories is used, however, because the important issue for personnel testing is the degree to which measurement is representative of "real world" situations. The categories chosen fell on a continuum ranging from artificial but highly controlled to realistic but uncontrolled situations. The higher the degree of control, the greater the loss of realism or representativeness of the research and of the measurement in it. Nevertheless, all measurement requires some degree of control or there is no standardization of measurement.

Laboratory Settings. This heading describes both actual laboratories, where full control of extraneous conditions can be maintained, and well-controlled simulations. Such control, in personnel testing, is rare except in experimental studies of human factors. Measurement in such research is usually concerned with the evaluation of a component of a system rather than the evaluation of a person or task as such. Individual proficiency in a complex skill, however, may be measured in laboratory-like simulations for certification purposes.

The emphasis is on the level of control rather than on the physical attributes of the setting. It is possible to have a highly controlled experimental study under carefully-selected field conditions. Measurement of certain attributes, such as physiological processes, may be done under conditions most nearly like those of laboratory control regardless of the physical setting in which they occur. Even within a laboratory setting, the level of control may vary; in the

study of reaction times, for example, a laboratory equipped with modern electronic apparatus can achieve a higher level of control, and therefore a greater degree of accuracy, than one where reactions are timed with a stopwatch.

The control referred to in this discussion is not experimental control over manipulations -- a major characteristic of an experiment -- but control over the measurement process itself. Without such control, attributes other than the one being measured (including attributes of different objects) are permitted to influence the measurement. With the highest levels of control, there is little influence on the obtained measurement from extraneous sources. For example, the electronic apparatus is more accurate in measuring reaction time because it does not include error due to the speed of reaction of the observer.

More accurate measurement is not necessarily better measurement. The basic problem in evaluating measurement under conditions of laboratory control is the problem of generalizability. Does measurement under the idealized, controlled conditions generalize to "real world" uncontrolled conditions? The question is an empirical one, and its importance varies with the opportunity for distortion in measurement in either artificial or clearly uncontrolled situations. What is at issue is the Brunswickian notion of representative design. Measurement taken under the relatively sterile conditions of laboratory settings may lack representativeness, and the laboratory may therefore introduce its own error by influencing the behavior or variable under study.

<u>Settings of Institutional Control</u>. This rather peculiar term is intended as an umbrella term covering employment offices, clinics,

training centers, and other settings in which meaurement is taken under standardized (if not really controlled) conditions — conditions which include the awareness of the subject being measured that institutional decisions are going to be based on the results. Standardization implies certain conventional concerns, such as consistency in time limits, instructions, formats, etc. There are other concerns, however, that have not been handled particularly well in the psychometric literature. For example, are testing conditions standardized when the same instructions are read to all people to be tested, or when all of the people to be tested have been brought to some common level of understanding? Answers to such questions may well determine the success in minimizing unwanted influence on measurements.

Field Settings. Realistic field situations can be described on several dimensions. One might be the number of constraints on performance imposed by the environment; in some environments one may perform a wider range of tasks, or perform them with more difference in quality, than in more constraining settings. Some settings are supportive and facilitate performance of the measurement task; others are hostile environments which make it difficult to perform well. Subjectively, environments fall along a continuum ranging from pleasant to unpleasant settings, or, alternatively, motivating as opposed to inhibiting conditions.

The purposes of measurement imply the kinds of real-life conditions to which the results are expected to generalize, and they also determine whether, under those conditions, one wants to infer maximum or typical performance. It is obvious that some situations place a limiting influence on performance; conditions of measurement may need to include similar influences. Other consequences of the setting include effects on performance standards or on what may be expected

as typical performance. Field settings, in short, provide numerous sources of influence on obtained measurements. These influences across settings may not be consistent from one individual to another; scores obtained in different settings need to be compared for means, variance, and correlations to determine whether inferences from scores generalize from one setting to another.

TYPES OF VARIABLES: ATTRIBUTES OF PEOPLE

Many kinds of variables are measured in psychological research, including attributes of organizational and physical climates, architectural variables, tangible objects, social relationships and many other stimuli or behavioral outcomes. For convenience, the discussion here will be restricted to attributes of people and to attributes of the tasks they are asked to do.

The infinite variety of attributes of people have been organized below in seven categories. The categories, which certainly are not exhaustive, seem less important than the order in which they are presented. The presentation begins with a class of variables most amenable to objective measurement and concludes with variables for which little or no objectivity in measurement can be claimed.

Objectivity in psychological measurement is an elusive concept. It certainly should not be, as is commonly done, confused with a multiple-choice format. The topic will be reexamined later. For the present, modifying an earlier discussion (Guion, 1965), three considerations may facilitate objectivity in measurement:

1. Objectivity is facilitated by responses which can be empirically verified against some external standard as opposed to qualitative or evaluative responses of unverifiable substance.

2. Objectivity is facilitated by responses which are free or unconstrained, where the respondent's own preferred alternatives may be expressed, as opposed to responses which are restricted or structured by the measurement process itself (Thurstone, 1948).

3. Objectivity is facilitated by responses not easily or likely to be distorted, as opposed to responses distorted by deliberate faking, anxiety about the purposes of the testing, etc.

The common element in these is a matter of inference. Inferences can be made with more confidence, and in fact are smaller inferences, if based on responses that can be declared accurate, or are free from format constraints, or are not distorted in other ways. On the other hand, inferences are shaky indeed from faked reports of internal states or from responses which fit the format but give the respondent no option for the response that would be a better, more accurate, or more honest response.

Physiological Processes. In personnel testing, physiological variables are rarely considered except in human factors or stress research. Nevertheless, it is instructive to consider how such variables can be measured. Examples might include such diverse variables as respiratory rate or capacity; pulse, blood pressure, or other cardiovascular measures; metabolic rates or chemical concentrations; visual, auditory, or cutaneous acuity or sensitivity; and others. Measures of such variables are often fundamental or mathematically formally derived measurements. They may be measured by counting or in physical units.

It is important to be clear about the variable being measured as distinguished from the variable that might be inferred from the measurement. If we are concerned about the effect of a program of exercise on cardiovascular functioning because the purpose of the

program is to improve cardiovascular functioning, for example, we measure such functions simply as variables to be interpreted on their own terms. Frequently, however, we may be interested in the same measurements as a basis for other kinds of inference. For example, research on reactions to stressful environments may measure the same cardiovascular functions for inferences about levels of anxiety, a distinctly different type of variable.

Motor Skills. This category, too, is concerned with biological functioning; it differs in that its variables are peripheral, usually directly observable behaviors; that is, the variables do not have to be inferred from readings of instruments. Variables included in this category include dexterities, coordination, strength, and other patterns of muscular behavior.

In personnel testing, these variables are most likely to be measured as predictors in selection systems or for research on safety. This category, and the preceding one, may on occasion be measured as aspects of work sample performance; a work sample test for firefighters, for example, may consist of timing the speed with which a candidate can climb a ladder and return. An inference is involved, but it is such an easy, direct one that it is not often questioned; it is easy to infer skill in doing something but it is inadvisable to infer a lack of skill from poor performance. The assumption is that one cannot perform well without skill, but lack of skill is only one of many reasons why one would perform poorly.

Performance Variables. This is an extremely broad category, including most overt behavior. It includes, but is not limited to all measures of proficiency, speed or quality of performance, evaluations of work products, ineffective or disruptive performance, or

certain kinds of performance habits or styles -- approaches to carrying out tasks. Such variables, whether defined in terms of maximum or of typical performance, are most often used in the role of criteria or dependent variables. They may also be used as predictors or as bases for instruments certification decisions. Measures of attributes of actual behavior may be the basis for certification of proficiency or acceptability, or level of proficiency may be inferred from measurements using indirect indicators. Work samples, in most cases, are examples of performance measures, but so also are the ubiquitous ratings by supervisors. Performance is usually an objective fact, but it does not necessarily follow that its attributes can be easily or objectively measured.

Job Knowledge. Closely related to the measurement of proficiency is the measurement of the knowledge required to become proficient. Often, although sometimes erroneously, job knowledge tests are used for drawing inferences of proficiency. This use of job knowledge variables needs to be recognized as an example of a formal hypothesis; that is, it is hypothesized that a measure of test proficiency is a function of measured job knowledge. The hypothesis may often be tenable, particularly in highly complex jobs, but it usually deserves an empirical test.

Of the categories so far mentioned, this is the first in which conventional principles and methods of test construction, following classical psychometric theory, are easily used. Psychometric principles are rarely considered in the techniques for measuring physiological processes. It is true that psychometric evaluations of reliability and validity are commonly applied to measures of dexterity and coordination, and they are frequently given lip service in measuring aspects of performance. Nevertheless, this category is the first in

the list in which there are individual items that can be clearly clustered into internally consistent dimensions, the kind of items for which classical theoretical propositions, such as the Spearman-Brown formula or the theoretical foundations for definitions of parallel test, were created.

Cognitive Variables. The history of mental measurement is largely a history of the measurement of cognitive processes. It began with the measurement of intelligence (or "genius"), and much of its progress has occurred through refinements in the methods of measuring intellectual functioning. Intellectual functioning is generally considered a form of information processing, the principal preoccupation of cognitive psychology.

Typically, cognitive variables in personnel practice are measured through the use of paper-and-pencil tests. In other areas of psychology, there is evidence of discontent with this form of measurement. Lunneborg (1977) reported a series of three studies using laboratory measures of reaction time correlated with standard paper-and-pencil tests. The correlations were rather low, but the attempt to understand conventional test performance in the language of cognitive processes seemed intriguing. Cognitive variables are among the most commonly used predictors in personnel selection and classification programs; attempts to measure individual differences in these variables that utilize cognitive theory and research should be watched with interest.

Aspects of Personality or Temperament. Attempts to measure characteristics of personality have been highly varied; they include personality inventories, projective procedures ranging from ink blots to sentence completion forms, and procedures for inferring personality

characteristics from such objective data as suggestibility during an experiment, etc. More recently, attempts to assess such variables through assessment center exercises have become quite popular.

These variables differ from those under all previous headings; they are less a matter of what a person can do than of what a person will do. The emphasis is motivational, and it has no objective referent. Characteristics of personality and temperament are therefore evaluated against normative standards.

Attitudes. The measurement of attitudes involves assessing affective reactions to a wide variety of environmental characteristics. Attitude scales may be developed by scaling checklist statements, writing single item questions with graphic rating scales or other ad hoc collections of intuitively scaled response options, or by using the method of summated ratings on a series of such questions or checklist statements. The most common example of attitude measurement in personnel testing is the measurement of job satisfaction and related reactions to work and work settings.

The level of blood sugar in a given sample of blood constitutes an empirically verifiable fact; there is no way in which the level of job satisfaction of an individual in a given setting can be considered similarly verifiable. Moreover, the methods of measurement of attitudes rarely permit free responses; the responses typically are constrained by one of the formats mentioned above. Moreover, as people try to interpret the purposes of the measurement, or fear that their responses can be identified and used against them, there is a strong probability that responses will be consciously distorted. In all respects, the measurement of attitude seems to be the least objective of any of the variables in this list.

TYPES OF VARIABLES: ATTRIBUTES OF TASKS

The long history of measuring attributes of people has made it possible to organize variables describing people in a fairly coherent way. There is no comparable history in the measurement of task characteristics, although the kinds of variables to be sampled in developing work samples makes task variables extremely important to the present paper. Very briefly, nine categories of task variables can be suggested. An attempt, tentative and faltering, has been made to suggest again a rough order of objectivity or verifiability, but no definition of objectivity is offered. The earlier treatment of objectivity in terms of responses is clearly not applicable.

Duration or Intensity of Attention. Some tasks, for example that of the air traffic controller, require a constant and unwavering vigilance for prolonged periods. Other tasks require less intense attention, and even that needs to be maintained for only brief periods. Variables might differ according to the sensory modalities involved, the focus of attention, or the nature and costs of the consequences of inattention. Some of these variables may relate more to cognitive than to sensory processes, such as the number or complexity of details that must be comprehended or manipulated, or the degree to which the task demands attention to fact as opposed to attention to broad generalization.

Hazards. Physical, social, or economic risks may be intrinsic components of certain tasks. Such variables need to be considered very carefully in the development of work sample measures; a work sample designed to assess the performance of a police officer in making an arrest may, for example, be severely distorted if the sample involves simulated conditions in which the officer knows there is no chance of being shot.

Degree of Task Structure. Perhaps one of the most widely studied attributes of tasks is the degree of uncertainty (or its opposite, structure). In some tasks the outcome of performance is highly predictable. That is, one knows very clearly that doing the task in one way leads surely to specified errors, whereas performance in a different way leads to acceptable work products. In contrast, other tasks, such as artistic or craft tasks, are often carried out with very little assurance that the result will be the one intended.

Organizational Involvement. Some tasks can be done in nearly total isolation. Other tasks require a worker to receive material or ideas from other people and may also influence work of other people; examples include assembly line activities, team activities, etc. Organizational involvement may be a single variable which can be measured in terms of the number of necessary interactions with other people in an organization required to perform a task satisfactorily; alternatively, it may be analyzed into component variables as different organizational entities as the locus of involvement.

Task Complexity. Variables under this heading include the level of knowledge and skill required to carry out the task, the variety of skills demanded, the number or complexity of choices or decisions that might have to be made, the level of accountability or damages in the case of inadequate performance, or even the learning time required to perform the task effectively. It is possible to develop a work sample test using performance on relatively simple tasks as a basis for inferences about performance on a more complex task. Doing so implies, again, a hypothesized relationship between performances on the simple and complex tasks, and that hypothesis needs to be tested before its tenability is assumed.

Intrinsic Feedback. On some tasks, a worker can obtain information about how well he is doing the task as he is doing it. One who is cutting a piece of wood or metal on a lathe, for example, can periodically check the dimensions against the specifications with calipers and can evaluate his work. If one is using expendable tools, such as saw blades, and one's rate of wear or breakage is excessive relative to some standard, he can be aware of the flaw in performance without being told by an independent observer or supervisor. In other tasks, feedback about quality of performance is long delayed and may sometimes be filtered through several processes; sometimes it comes only from the subjective judgments of peers or supervisors. Work sample testing appears to be more easily directed toward tasks with opportunities for some intrinsic feedback.

One set of feedback variables may relate to the size of the task unit. The amount of time or number of cycles required to complete a unit of work, the frequency of interrupted tasks, the opportunities to set goals, the tempo or pace of the work -- all of these influence the degree of feedback one gets in performing tasks. (For a discussion of these variables, see Ryan & Smith, 1954.) Once again, the importance of such variables in psychological measurement by work sample tests is that work sample tasks should have feedback properties similar to those of the work being sampled.

Skill Demands. This category includes motor, sensory, and cognitive skills (and perhaps even attitudes) that are clearly prerequisite to effective task performance. For some of these variables the task may demand quite high levels; for other variables, the level of ability demanded by the task may be much lower. These variables have special implications for work sample testing to whatever extent they change over time. Changes in the skill demands

of the job may correlate with, but should not be confused with changes in the skills of a person doing the job (Alvares & Hulin, 1973). Some changes in skills applied in the performance of a job occur with accumulated learning through experience; if this happens, the advisability of work sample testing of inexperienced people should be questioned.

Significance. This category is intended to include any variables which evaluate the importance of task outcomes. It may include the importance of the task as an influence on the performance or satisfaction of other people within the organization, it may be an element of importance for society at large, or it may involve importance for client or customers of the organization.

Autonomy. Some tasks can be performed by the worker without supervision or advice from other people; others must be done with close supervision or consultation. Autonomy is the degree to which the worker is free to do the task without the permission or advice of someone else. Another kind of autonomy might be defined as the worker's degree of discretion in making decisions; there may be different levels of discretion for different kinds of decisions about the way tasks are to be performed or the sequence to be followed in performing them. Or, autonomy might be the number of tasks that can be completed, or the period of time one may continue to work, without seeking authorization. Or, it might be the level of the worker's control over such things as pace, or sequence of activities, or quality or quality standards.

TYPES OF MEASUREMENT METHODS

The methods for measuring task attributes are related to those

for measuring the attributes of people; what differs is the nature of the inference drawn. Although a variable such as the degree of physical hazard may be determined by counting accidents, it is more often assessed by someone's judgment or perception -- a cognitive process of the observer.

All measurement of psychological attributes begins with the observation of the responses people make to specific stimulation. Differentiation among measurement techniques is necessarily based on the nature of the observational aids used and on the manner of recording responses and transforming them into measurement.

Five categories are listed. Once again, these categories are listed in the order in which they permit objectivity in measurement or, conversely, in the reverse order of the magnitude of inferential leaps necessary for the evaluation or interpretation of data. Again, as before, the categories follow this order as a matter of convenience, not as a matter of invariance.

Instrumentation. Instrumentation as used here refers to equipment, such as mechanical, electronic, or optical aids for observation. People may respond to an emotional stimulus with an increase in the moisture content of the skin surface. Except in the strongest emotional states, however, these increases may be imperceptible without the aid of galvanometers.

Many physiological responses are measured on standard polygraph instruments. Most psychological laboratories boast an array of solid state electronic circuitry for the measurement of reaction time that would have seemed like science fiction to the psychologist holding a stopwatch a mere quarter of a century ago. Sophistication in

research and sophistication in instrumentation have developed in tandem.

Instrumentation is commonplace in measuring sensory capacities or reaction times or choices. For the former, it is especially helpful in the presentation of stimulus materials, while the latter uses instrumentation to magnify, clarify, count, or record responses or characteristics of responses. The instruments may be highly sophisticated or quite simple. They may often be developed specifically for particular measurement problems. For example, Gessewein & Corrao (1971) developed special apparatus to study the possibilities of leg fractures. Their purpose was to develop a family of curves to provide designers with the means of predicting those conditions under which Naval personnel on ships would be likely to receive fractures; the variable to be measured was the force of impact as a person fell from various heights, and the technique of measurement was to have subjects drop stiff-legged onto a force gauge platform.

Instrumentation is often used in inferring work sample proficiency through measuring characteristics of the work product. In a work sample requiring the subject to make solder connections, for example, the quality of response might well be measured by measuring the conductivity of the solder connections themselves rather than by measuring responses directly. If a piece of metal is to be machined to specifications, the resulting product can be measured with anything from a ruler to laser beams to determine whether the product is within tolerances.

Direct Observation and Recording. This category is best illustrated by research in applied behavioral analysis which requires observers to count frequencies of specified behaviors. Just as

measurement techniques with instrumentation vary greatly in sophistication, so also measurement by direct observation varies greatly in the clarity, detail, and precision of instructions to observers and in the precision with which their observations may be recorded. Under many circumstances, some form of instrumentation may be a portion of the recording process. That is, the observer may make frequency counts either by making tally marks on a piece of paper or by pressing a button activating a counter.

A less exact form of measurement by observation is used in many assessment center exercises. The observers may have no specific behaviors to count; instead, they may be instructed to observe and write down "any salient behavior." At the conclusion of the exercise, the observer's record may consist both of such narrative descriptions and an evaluative rating of the behavior observed.

Records and Biographical Data. Many variables, of which attendance is perhaps the best example, are measured by frequency counts obtained not by direct observation but by examination of recorded data. Many kinds of records are maintained in most organizations. If they are maintained consistently and accurately, they provide useful data sources for the development of a variety of measures. Therein, of course, lies the rub; most systems of personnel accounting are notoriously poor. It is, however, possible to develop and maintain effective ad hoc record systems for periods of perhaps several months.

Measures of many kinds of variables may be derived from data maintained in records. For example, records may contain frequency counts of production and may also indicate periods of time away from the principal assignment when a worker cannot be expected to be

productive. By combining the two sets of data, derived measures of productivity per hour or per day can be developed. If situational factors influence daily average productivity, records can be organized so that distributions of productivity in different situations can be determined with individual production records standardized in terms of those distributions.

Records are kept in memory banks, be they file drawers, computers, or human memories. If the memory bank is in a computer, it is simply a form of storage. However, data stores in the memory of an individual is often changed in "storage" and retrieval processes. Many variables are measured by asking individuals to pull from the records of their own memories information which can be scaled, counted, or classified. It is in this context that the major difficulty in such measurement comes into clear focus: the accuracy of records must always be suspect. Records, whether from the memory of individuals or from files, suffer from variations in carefulness, in organizational procuedures, in the interpretations of numbers, and in many other ways that distort their accuracy.

Testing. Personal attributes of people are most often measured by asking them questions and recording the answers to those questions; this is certainly the most common measurement technique in personnel research. Sometimes the questions are actually assignments ("Solve this problem" or "Assemble that gadget"), but the prototype of this form of measurement is the multiple-choice test item. The stimulus material is the question asked or implied in the stem, and the response is the choice of the option considered correct. If the item has a genuinely correct answer, as in an arithmetic problem, the correctness of response is highly verifiable and such tests are usually called objective. There is less verifiability of the

correctness of the response when the question deals with the subject's own typical behavior. A question might, for example, ask the subject how he prefers to spend his spare time. The optional answers might include responses such as reading a good book, going to an art museum, attending a symphony orchestra concert, or watching situational comedies on television. Many people will, of course, literally spend more time watching situation comedies if for no reason other than the ready availability of a television set; symphonies, art museums, and good books may not be as accessible. The question, of course, does not ask a factual question of how one's time is literally spent; it asks how the subject likes to spend his time, and the response to that question is not at all verifiable. Only the subject himself knows his own preferences, and he may not be sure of them. Even if he is sure, he may not be truthful. If he actually prefers situation comedies over concerts, he may nevertheless respond that he would prefer to go to a concert simply because in the testing situation he perceives this to be a more socially desirable response. Since there is no direct way to determine whether an individual has responded honestly to the question, or even whether there is a clear-cut answer, such testing is considered highly subjective.

Although the written multiple-choice question is a prototype, it is by no means the only approach to measurement by question and answer techniques. In determining how well an individual might be able to detect salient stimuli in the midst of irrelevant but pervasive stimulation, the question might be, "In which quadrant is the target stimulus?" referring to a projection on a screen. Questions in any form must be phrased appropriately. In the familiar Snellen Eye Chart, for example, the "question" may be, "Can you read the next line?" It is not appropriate for the subject to answer with a yes or no; such flippancy can be avoided by simply assigning the

reading as a task: "Now read the next line."

Ratings. When all else fails, or when energy or imagination is lacking to suggest anything better, psychological measurement consists of ratings. Some form of rating (or, more generally, subjective evaluation) is the most commonly used method of measuring performance and related behavioral variables. The basic rating system consists of a format for recording subjective evaluations of designated stimulus objects or items; the familiar graphic rating scale is only one example.

In fact, better examples involve both descriptions of observations as a basis of evaluation and the evaluation itself. The observer may note behaviors and either rate the behaviors along some designated scale or consider them in rating the ratee on a pre-determined dimension. Occasionally, the observations themselves form a rating scale. Much research in developmental psychology or in animal research requires observers to check one descriptive behavior statement observed among a list of behavior statements that have been previously scaled.

Ratings are often not based on systematic observations. Periodic efficiency reports or other methods of performance evaluation frequently consist of ratings based on the vague impressions of superiors who may never have had an opportunity to observe the subordinate's behavior directly. Research on this ubiquitous use of ratings casts considerable doubt on their utility.

Serious question may also be directed to the many forms of self-rating used in psychological measurement. Many personality inventories of a question-and-answer form require that answer to be given in terms of a scaled response. An item describing a particular form of behavior might, for example, call for response options scaled in four

steps: "very much like me," "somewhat like me," "not very much like me," "not at all like me." This, too, is a subjective judgment in which the response requires a rating along a scale. Subjects may often be given simply the assigned task to rate themselves on specific dimensions -- again with the rating to be placed on a designated form.

The objectivity of ratings, or their verifiability, depends primarily on the nature of the stimulus material. Subjective ratings, or discriminations, are called for in any psychophysical measurement, such as an eye examination, yet these may be treated as relatively objective. In contrast, an instruction to rate someone on "quality of performance" is far too ambiguous to permit an interpretation of objectivity. Moreover, the objectivity of ratings depends largely on the raters' desire for objectivity; many forms of bias, ranging from the self-protection of a central tendency response bias to overt prejudice may influence recorded ratings.

IMPLICATIONS OF THE CLASSIFICATIONS

The classification schemes described in the preceding section may prove unwieldy or ambiguous if they were used to classify actual studies; it has not been empirically tried. A desirable next step would be to ask different expert judges independently to fit real examples into the categories described. If specific uses can be classified easily and reliably, support for the taxonomy would be inferred; unreliability in classification would identify needs for modification.

For the present purposes, however, no tightening of the taxonomy is necessary. These categories may not be optimal, but they are at least indicative; their implications for the construction and evaluation

of new testing programs will not differ substantially from those of an empirically modified scheme.

In this section of the report, implications will be considered first for each of the different classification schemes; they will then be considered for combinations of classifications.

IMPLICATIONS OF PURPOSES

1. For all purposes, measurement leads to decisions, and these in turn at least imply some prediction of outcomes of the decisions.

2. Work samples may be relevant for any purpose, either as dependent variables or as independent variables.

3. No class of purposes imposes restrictions to particular kinds of measurement. Although measurement of some aspect of performance is commonly intended for many of these purposes, it can be based either on fundamental descriptive measurement or on measurement requiring greater inferential leaps. Measurement in program evaluation for organizational decisions, or measurement calling for the certification of proficiencies, should in general need smaller or easier inferences than do measurements for other purposes.

4. The different purposes impose no special restrictions on the kinds of variables to be assessed; both task variables and person variables need to be assessed in meeting many of these purposes.

5. Measurement techniques which maximize variance may be used for any of the types of purposes and are highly to be desired for most.

6. Measurements taken for decisions about groups (primarily in evaluations of material, processes or groups, but sometimes in organizational trouble shooting), should provide significant group differentiation. The principle may also apply to certification (for example, to differentiate masters from nonmasters), but only if the groups are very carefully defined and if the basis for group membership is stable. These two conditions may often be impossible to satisfy.

7. Where the purpose is prediction, the evaluation of measurement must be based on how well the predictor measure correlates with a measure of the future event or state to be predicted.

8. For diagnostic or certification purposes, measurement should be evaluated by logical or statistical relationships with broader indices of proficiency or the diagnostic categories. Such evaluations can be based on the logic of content sampling, on correlations, or on experimental results.

IMPLICATIONS OF SETTINGS

1. The purposes of measurement define the set of conditions most appropriate to that measurement; this set of conditions might be termed the target conditions. In any setting differing from the target conditions, the measurement setting should be representative of the target situation in salient respects.

2. Different settings may be responsible for different contaminating variables in measurement; interpretations of the results of measurement should consider the possible distortions introduced by a particular setting.

3. Where the measurement situation differs significantly from the target situation, the generalizability of inferences from the one to the other must be assessed.

4. Measurements in laboratory settings or simulations may fail to generalize if they are over-controlled, that is, if influences expected in the target situation are not permitted to vary in the laboratory.

5. Generalizability of measurement in institutional settings is less concerned with the generalizability of scores than with the generalizability to attributes of greater institutional concern; usually, this form of generalizability is expressed as predictability.

6. When measurement is done in naturalistic or field settings, standardization requires that specific sets of conditions be used. The problem of generalizability is, in such settings, one of generalizing scores (or inferences from scores) obtained in the standard setting to other relevant settings.

IMPLICATIONS OF PERSONAL VARIABLES

1. The variables in the higher categories on this list are more likely to be tangible or directly observable and less likely to be abstract. Therefore, they can be measured more objectively, and mathematically formal methods of measurement are more likely to be available.

2. The higher the category on this list, the less appropriate is conventional norm-referenced measurement. One's pulse rate after a period of extensive exercise is not evaluated by its position in a normal distribution of pulse rates; it is evaluated with reference to a standard given the age and exercising condition of the individual whose pulse is measured.

3. Variables high in this list are likely to be evaluated primarily in terms of accuracy; accuracy is an irrelevant concern for variables low on the list. The notion of accuracy implies a well-calibrated scale of measurement, usually in units accepted by the scientific community.

4. Work sample tests are most likely to be developed to measure aspects of task performance, although in some components and under some circumstances they may measure job knowledge variables, motor skills, or physiological processes. Since work sample testing measures variables in the higher categories, these variables should be objectively measured, interpretable with reference to a priori standards, and capable of accurate measurement on a well-calibrated scale.

5. The literal measurement of one variable (e.g., skin resistance to current) may be chosen as a basis for inferences about a different variable (in the example, it might be anxiety). Such inferences imply hypotheses that need empirical verification if the inferences are to be considered valid.

IMPLICATIONS OF TASK VARIABLES

1. The variables higher on the list, in general, are associated with greater opportunity and need for objective measurement; they should be interpretable with reference to previously established standards and accuracy of measurement.

2. The identification of classes of task variables helps to define the nature of a work sample test; a first stage (and sometimes sufficient) step in the evaluation of such a test is to evaluate the degree to which it is congruent with the work being sampled on salient classes of variables.

3. Performance variables in the list of personal variables are likely to be influenced both by task variables and by settings.

4. The overall nature of a task changes with changes in settings; it follows that a major consideration in measurement of task variables is the generalizability of scores or of inferences. As a specific example, the task of cleaning a rifle in the quiet of a barracks is quite different from the task of cleaning the same rifle, with the same dirt, under fire. If a task is to be properly sampled in a work sample, the conditions of performance to be inferred should be specified. Whether performance of the task under conditions other than those specified will generalize to those conditions is an empirical question.

IMPLICATIONS OF MEASUREMENT METHODS

1. The greater the precision in specifying the response to be observed, the less the ambiguity and the greater the objectivity of measurement. Methods higher on the list promote greater specificity.

2. The more objective or fundamental the measurement technique (for example, counting frequencies), the less the inference required. Of course, one may use a fundamental measurement for an intuitive inferential jump from it; such inferences usually need empirical verification. In general, inferences based on methods high on the list are more easily verified than those based on methods low on the list.

3. Regardless of measurement technique, some form of reliability information is essential to measurement. That reliability may be the consistency assured by well-calibrated instruments, or the agreement of independent observers, or the internal consistency of scaled responses to a set of attitude items. Whatever the form of reliability of greatest concern, no measurement technique can be evaluated more

highly than the reliability permits. Reliability is rarely a sufficient evaluation, even though it is a necessary one. A set of ratings may be highly reliable because of the presence of constant errors, but the reliability is of very little value if it means no more than consistently false inferences.

4. Objectivity may be illusory. The presence of sophisticated instrumentation is not an assurance of objective measurement. The question must be asked whether the measurement obtained with such instrumentation is fundamental measurement, that is, measurement to be interpreted in terms of its own units, or whether it is a basis for a derived inference.

SIMULTANEOUS IMPLICATIONS OF VARIABLES AND METHODS

Special implications for the evaluation of measurement can come from a simultaneous consideration of the kinds of personal variables being measured and the method of measurement. In abbreviated form, condensing the classification of person attributes to five categories, the two classifications are shown in matrix form in Figure 1. The matrix is so arranged that the upper left-hand corner represents the maximum possibilities for objective measurement and the lower right-hand corner represents the maximum in necessary subjectivity.

In the extreme cases, measurement of physiological or psychomotor attributes with special measuring instruments requires only accuracy in the calibration of the measuring instruments; with accuracy, questions of reliability are moot. Concern for the generalizability of measures obtained from the situation of actual measurement to targeted situations is, of course, always a consideration in the evaluation of any measurement, but so far as the variables and methods are concerned, the closer the situation to the upper left of Figure 1, the more salient the concept of accuracy is to the evaluation of measurement. Accurate measurements are those that are most readily verifiable with

	Instrumentation	Direct Observation	Testing	Records	Ratings
Physiological or Motor	B,D	A,B,D	A,B,(D)	B,D	A,C
Performance (Behavioral)	B,D	A,B,C,D	A,B,D	A,B,C,D	A,C
Job Knowledge	B,C,D	A,B,C,(D)	A,B,(D)	A,B,C,(D)	A,C
Cognitive or Personality Constructs	(A),C	A,C	A,C	A,C	A,C
Attitudes	(A),C	A,C	A,C	A,C	A,C

Figure 1. Implications for type of variable and type of measurement technique for the evaluation of measurement. A refers to reliability, B to acceptance of the operations, C to acceptance of inferences beyond the content, and D to interpretations relative to a standard.

reference to some standard unit of measurement such as an inch, a gram, or a count.

In the other extreme is measurement using some form of rating scale for the assessment of attitudes. There is no way in which the "accuracy" of such measurement can be verified. It is possible to obtain indices of consistency of response, but there is no way to determine whether the attitude is correctly or accurately measured. Not only are there no standard units of measurement, but there is no external referent that can be clearly said to be a better or more nearly precise statement of attitude; there is no Bureau of Standards for attitude measurement. Not even behavioral observations can be used as criteria for validating a measure of attitude; too many learned variables influence the expression or inhibition of behavior appropriate to the attitude. In a taste preference study, for example, one must simply take the subject's word for it that he evaluates one stimulus higher than the other. Thus the first kind of implication for this matrix is its influence on the permissible precision of measurement.

The above comments demonstrate an interdependence of the nature of the variable being measured and the method of measurement. Both the nature of the variable and the nature of the technique influence the saliency of different considerations in the evaluation of measurement.

Reliability. Beyond generalizability, which is universally necessary, the various cells in Figure 1 identify up to four kinds of essential evaluations for particular combinations. Cells marked with an A are those in which the first step in evaluation is an inquiry into reliability. The first step in evaluating reliability is not a computation of a reliability coefficient but an examination of the

technique of measurement itself: is the method of measurement appropriately standardized? Beyond that, the question of reliability encompasses all of the familiar concerns of equivalence, stability, and, above all, internal consistency.

In a sense, every cell in the matrix should include an A since reliability is the sine qua non of effective measurement. The cells of the upper left-hand corner, however, will have satisfied the needs for reliability automatically if the measurement can be shown to be accurate. Since accuracy has been identified as the principal consideration for this set of combinations, and since unreliable measure cannot be very accurate, then the evaluation of reliability is superfluous if accuracy is established. In all other cells, reliability often must be established as a basis for, or at least a consideration in, any other evaluative determination. Where special instruments are used, reliability may refer primarily to technical fallibility (such as trouble from poor electrical contact). Where measurement uses observers, the consistency or agreement among observers is the essential reliability. In some forms of physical or behavioral observation, the observing and recording responses may be easy enough that little or no observer error is possible or likely, and it may in such cases be unnecessary to become greatly concerned about reliability. Where observers are rating knowledge or cognition or attitudes, they are exercising their own judgments and, therefore, the likelihood of fallibility in measurement because of differences in observer judgment is very real and must be investigated. Reliability in measurement by testing is well-established in classical psychometric theory, as it is in scaling and other forms of rating. Reliability in record keeping is probably derivable from psychometric reliability; the consistency of record keeping, as well as the consistency of inferences, may be best determined by dividing records into small units of time

and comparing the data collected in different time periods. The various considerations needed for estimates of reliability will be reconsidered in the discussion of generalizability.

Reliability, it must be emphasized, is necessary in all measurement. It does not follow from that fact that reliability coefficients must always be computed. Where there is evidence of accurate measurement, it is also evidence of reliability, because there is no accuracy without reliability. Likewise, where there is evidence of validity (discussed below as "acceptability of inferences"), it is also evidence of reliability, because there is no validity without reliability. The important thing is to build the measuring instrument with care to insure maximum reliability.

A notation (A) in Figure 1 denotes particular uncertainty about effective ways to estimate reliability.

Logical Acceptability. Once reliability is established, the next evaluation concerns the acceptability of the operational definition, shown as B in Figure 1. This is largely a matter of precision in measurement; if measurement is fundamental in nature, following formal mathematical axioms, acceptance is highly probable. Statistically derived or intuitive measurements may also, however, be widely accepted, simply on the basis of the way in which the measurements are collected, if their logical foundation is persuasive enough. One issue in determining logical acceptability is whether the measurement fits its purposes in relation to the distinction between maximum and typical performance. If the purpose of measurement is to find out what people actually do in real situations, a highly controlled estimate of maximum performance cannot be accepted on logical grounds, whereas a less sophisticated form of measurement obtained under more realistic

conditions -- i.e., more representative conditions -- may be readily accepted.

The greater the objectivity in measurement, the greater the likelihood of its logical acceptability. Objectivity, it should be noted, is clearly distinguishable from construct validity, despite points of similarity. As defined in this report, objectivity depends on the degree to which the response itself is free from distortion, whereas construct validity refers to the degree to which the interpretations from the response are free from distortion by influences unrelated to a designated construct. Probably the greater the objectivity, the greater the construct validity, but the question really does not arise. What does arise is the question of whether the response is a clearly identifiable, interpretable, unambiguous response as opposed to the degree to which it is undefined and subject to varying interpretations. An inference, even from some physiological measurement, may lack construct validity even when variables are accurately measured. In medical diagnosis, for example, physicians may find symptoms easily measurable but difficult to interpret diagnostically.

Under certain circumstances, characteristics of distrubutions of measurements may be considered in evaluating the logical acceptability of measurement. As just one example, one may ask whether the measurement involves ceiling effects such that descriptions of individuals high on a given attribute are inaccurately obtained because of the inadequacies of the measuring technique.

Perhaps the greatest boost to the logical acceptability of a measure (well, at least its acceptability) is what is known as <u>face validity</u>. The term is unfairly maligned, simply because it does not in fact describe an aspect of "real" validity. Nevertheless, face

validity is of both practical and technical importance. It is practically important because it facilitates judgments of logical acceptability, at least in the middle set of cells in Figure 1. It is technically important because examinees or observers may be more appropriately motivated by measures that "look right," thus adding to the objectivity of measurement.

Acceptability of Inferences. Another set of questions refers to the acceptability of inferences extending beyond the obvious content. In the conventional way of talking about psychometric validity, most of the preceding discussion on logical acceptability referred to so-called content validity. Questions of the acceptability of inferences are, in contrast, questions of construct or of criterion-related validity. The cells marked C in Figure 1 are those where attributes can be satisfactorily inferred from the measurement only on the basis of supporting empirical evidence. In any specific case, if the nature of the measurement is inference rather than fundamental description, the psychometric concepts of the validity of the inferences are the most important aspects of evaluation. Even if the measurement ostensibly measures at a more fundamental level, inferential jumps from that level must be validated. The example given earlier should be remembered: when one uses a physiological measure not as a description of physiological functioning but as a manifestation of anxiety, the inference to be validated is the use of the measurement as an index of anxiety. The accuracy of measuring the physiological process is irrelevant. Wherever the measurement is intended to lead to an inference of attributes outside of its literal content, evidence of some form of validity, specifically criterion-related or construct validity, is essential.

The crux of classical psychometric validity is the extent to

which the variance in measurements is attributable only to the construct intended to be inferred. Insufficient validity, therefore, means that part of the variance in a set of scores is classically seen (a) as being attributable to sources of variation other than the one intended or (b) as irrelevant to the variable to be predicted.

Standard-Based Interpretations. The letter D appears in Figure 1 wherever the obtained measure should be interpretable with reference to a standard. (Some arguable cells are identified with the D in parentheses. These are generally conditions permitting substantial objectivity and in which the accuracy of measurement can be assessed. Usually, they are examples where fundamental or mathematically formal measurement is plausible.

In a sense, this could apply to all of the cells; arbitrary standards or cutting scores could be established. Cognitive test scores, for example, can be interpreted as deviations from such arbitrary points.

The intent of the designation in Figure 1, however, is somewhat different; it is intended to refer to standards defined in terms of the measurement scale, not distribution of measurements. The intent here is not so much permissive as suggestive. Wherever the purpose of measurement is certification or institutional decision-making, the aim of test specialists should be to provide measurement that can be interpreted with reference to real performance standards.

SUMMARY

Heuristic classifications of the purposes and circumstances of psychological measurement, of the variables to be measured, and of the

techniques available for such measurement have been presented. Two major conclusions should be drawn. First, conventional psychological testing is contained in only a relatively small portion of all of the classes of psychological measurement. A single-minded devotion to the principles and theory of classical psychometrics has many values, but it also has the severe disadvantage of ignoring the values of other approaches to measurement. Other approaches may be more useful where accurate descriptions rather than abstract inferences are sought; even testing for inferential purposes can be improved if the methods of obtaining the underlying descriptions are more objective and accurate.

Second, classical psychometric theory may be too narrow to use in the evaluation of measurement in some of the classes. Evaluation of measurement may include reliability and validity estimation, to be sure, but it should also include a logical evaluation of measuring techniques as operational definitions of variables, and it should seek more frequent application of the usual scientific practice of interpreting measures with reference to a priori standards.

The classifications, and the broad conclusions reached from considering them, apply to work sample testing. Work samples may be used for any of the purposes of measurement, although in these reports they are primarily considered for certification purposes. Whether the product is scored or the process of getting it, work samples fit in any kind of setting; again, however, the interest of this report is primarily in settings of institutional control. With reference to Figure 1, work samples are most likely to be tests of performance, although they may include any of the classes of variables represented by the top three rows or the classes of methods in the three columns on the left.

The common requirements for the evaluation of measurement in those nine cells are (a) assessing the logical acceptability of the measurement and (b) the possibility of interpreting scores with reference to a standard. Neither of these kinds of evaluation invokes classical concepts of validity, although evidence of validity may provide further argument in the logic supporting a measure as an operational definition of the variables measured. Moreover, conventional validity is probably necessary for job knowledge or for some performance variables if these are assessed by direct observation instead of through tests or physical instrumentation. In short, despite the fact that conventional validities may provide useful information, inferences of attributes beyond the obvious content of the work sample itself are often conspicuously absent from work sample testing and, for these cases, conventional statements of validity may be superfluous and even misleading.

This is not meant to imply that criterion-related or construct validation of inferences from work sample performance is necessarily inappropriate. The point being stressed here is that the evaluation of work sample measurement is not fundamentally an evaluation of its use in the measurement of an inferred construct or of its power to predict some external behavior; rather, a work sample is evaluated primarily on its acceptability as a direct description of the performance of interest. The demands of this kind of evaluation need careful explication.

REFERENCES

Boyd, J. B. Interests of engineers related to turnover, selection, and management. Journal of Applied Psychology, 1961, 45, 143-149.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.

Dobbins, D. A., & Kindrick, C. M. Jungle vision V: Evaluation of three types of lenses as aids to personnel detection in a semideciduous tropical forest. Army Tropic Test Center (Canal Zone), Report No. RR-5, December, 1965.

Gessewin, J., & Corrao, P. Testing and modeling standing man's response to impact with applications toward predicting leg fracture to shipboard personnel. Washington, D.C.: Naval Ship Research and Development Center, Report No. NSRDC-3656, June, 1971.

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), Psychological principles in system development. New York: Holt, Rinehart, & Winston, 1962.

Guion, R. M. Personnel testing. New York: McGraw-Hill, 1965.

Lazarsfeld, P. F. The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer et al. Measurement and prediction. New York: Wiley, 1950.

Lord, F. M. A theory of test scores. Psychometric Monograph No. 7, 1952.

Lunneborg, C. E. Choice reaction time: What role in ability measurement? Applied Psychological Measurement, 1977, 1, 309-330.

Maier, M. H., Young, D. L., & Hirshfeld, S. F. Implementing the skill qualification testing system. U.S. Army Research Institute, R&D Utilization Report 76-1, April, 1976.

Melton, A. W. (Ed.), Apparatus tests. AAF Aviation Psychology Report No. 4. Washington, D.C.: Government Printing Office, 1947.

Muir, D. E. A critique of classical test theory. Psychological Reports, 1977, 40, 383-386.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Ryan, T. A., & Smith, P. C. Principles of industrial psychology. New York: Ronald, 1954.

Shimberg, B., Esser, B. F., & Kruger, D. H. Occupational licensing: Practices and policies. Washington, D.C.: Public Affairs Press, 1972.

Thurstone, L. L. The Rorschach in psychological science. Journal of Abnormal and Social Psychology, 1948, 43, 471-475.

Thurstone, L. L. The measurement of values. Chicago: University of Chicago Press, 1959.

Torgerson, W. S. Theory and methods of scaling. New York: Wiley, 1958.